PRML Assignment 4

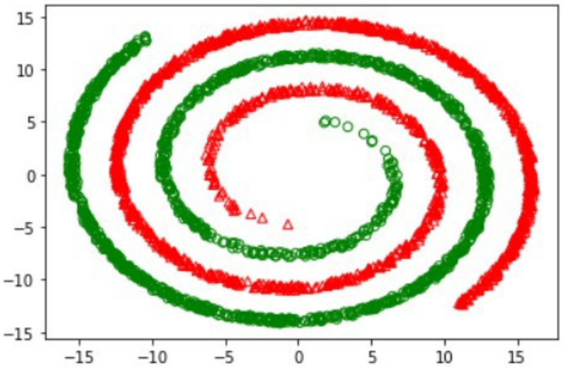
Rongali Rohith (EE19B114) Santosh G (EE19B055)

I. Aim

- To implement KNN, Logistic Regression, ANN and SVM on 4 types of data sets, namely, Synthetic Data set, Image data set, Handwritten Data set, Spoken digit data set. To implement PCA and LDA to transform the data; and to repeat the implementation again.

II. Synthetic Data

- The data belongs to two categories, which can separated by a non-linear decision boundary.
  The following is the initial data given, the data has been plotted to understand the complexity of the data and possibles issues with the models to be applied.



A. K - Nearest Neighbours (KNN's):

- The test data consists of points plotted on the plane (as the data is 2-dimensional).
  Now for each point in the Development Data, K-Nearest neighbours are computed based on the Euclidean Distance and based on the majority, the label for the dev-data is given.
- We have checked for various values of K, and we have achieved an accuracy of 99.9% for values of k even as low as 5.
  Confusion matrix for KNN (K = 5):

	Belongs to Class 1	Belongs to Class 2
Classified as Class 1	500	0
Classified as Class 2	1	499

B. Logistic Regression

- We have implemented Logistic Regression on the data to compute the results but as the train data is highly non-linear, the efficiency was poor and only resulted in an accuracy of 65.9%.

C. Support Vector Machine (SVM):

- SVMs work the best, when the data is linearly separable; but the data is not linearly separable in lower dimension (given 2-dimension).
- By Cover's theorem, the data can be made linearly separable in higher dimension and hence we use the "RBF Kernel" to achieve better results from the SVM Model.
- Once the data is transformed into linearly separable form, the SVM creates a decision boundary and based on that decision boundary, the points in development data are classified.
  SVM resulted in the same accuracy as the KNN Model, 999 of 1000 have been classified correctly and only 1 has been mis-classified. The confusion matrix is same as above.

D. Artificial Neural Network (ANN)

- Artificial Neural Network consists of various layers consisting of nodes that converge as the layers proceed.
- Each layer can have different activation functions such as ”Relu”, ”Sigmoid”, ”Softmax” etc
- We feed the input into the input layer that is further fed into nodes after various computations.
- To get the final output, the final output layer has been activated using the softmax function, hence the outputs will be in probability for each class.
  As we are also supposed to provide the labels, it is clear that ANN is a supervised model.
- We have used 4-Layers with 64, 16, 4, 2 nodes in each layer, as there are two classes, they have been gray-coded as 01 and 10, now the output-layer with 2-nodes will either give an output of 01 or 10 conveying the estimated state. This model has resulted in 100% accuracy and the ROC has been plotted below:

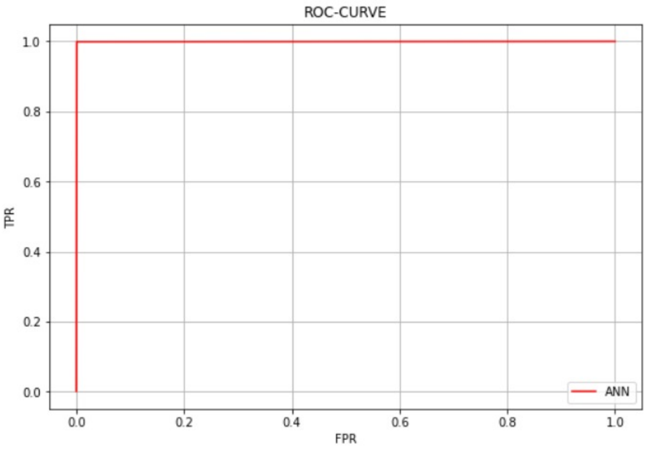


Fig. 1. ROC for ANN on Synthetic Data

- As the accuracy is about 99.9% and more, the ROC curves for each model is almost the same and the DET Curve is almost empty.

E. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)

- The following are the plots of the two principal components and LDA vector. Clearly, projection on to any of the principal components or the LDA vector will not help us as there will be quite a lot of overlap between the two classes.

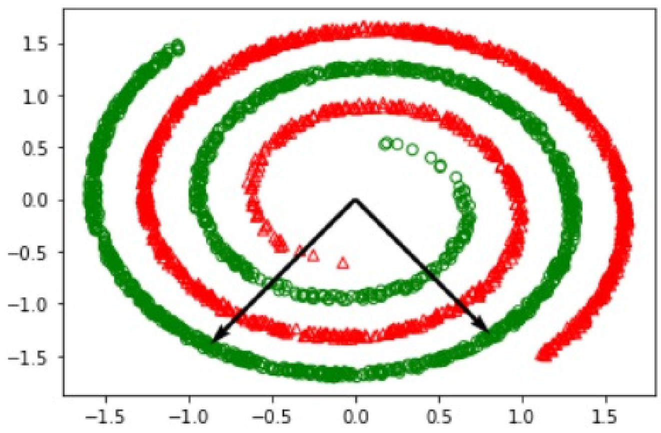


Fig. 2. PCA Compontents of Synthetic Data

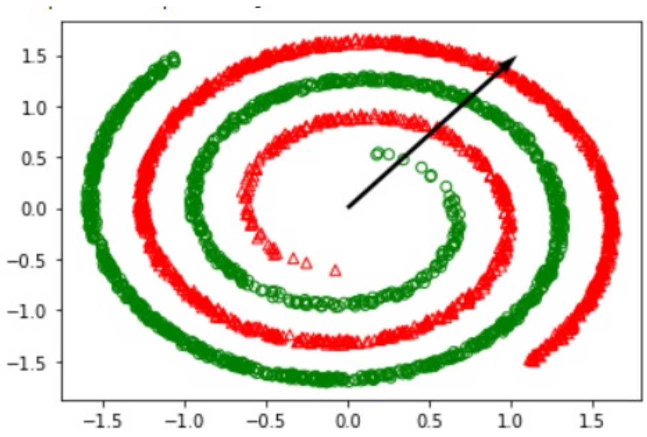


Fig. 3. LDA Vector of Synthetic Data

III. Image Data set

Image data set contains images belonging to 5 different classes. Each Image is broken into 36 blocks. Every image has 23 dimensional feature vectors. They are flattened to 828 sized vector, and we shall apply the classifiers on these vectors.

We train the four classifiers on the given data set and observe the following results:

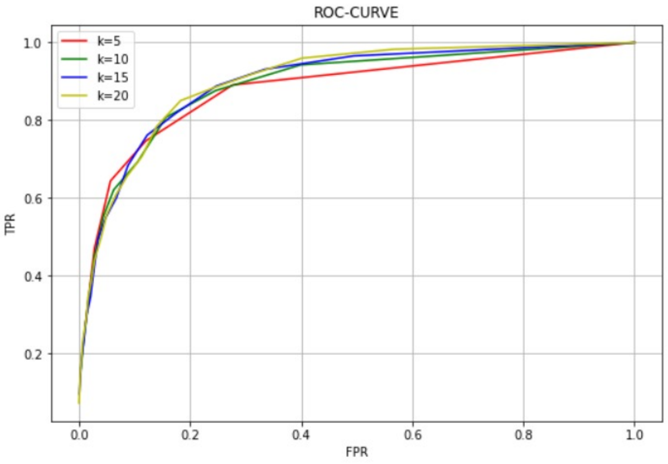


Fig. 4. ROC Curve of KNN applied on Image data

We have experimented the same with various values of K, the number of nearest points chosen; but the accuracy remained almost the same in each case.
Now we transform the train data by applying PCA and LDA; we can observe increase in efficiency in case of PCA but the efficiency reduced when LDA is applied,
The above can be explained by the following: BY performing PCA we are removing redundancy, but with LDA we are restricting the dimension to 4, hence losing the data and accuracy.
It is also to be noted that the data has to be Gaussian and linearly separable for LDA to perform well; which might not be the case in Image Test Data.

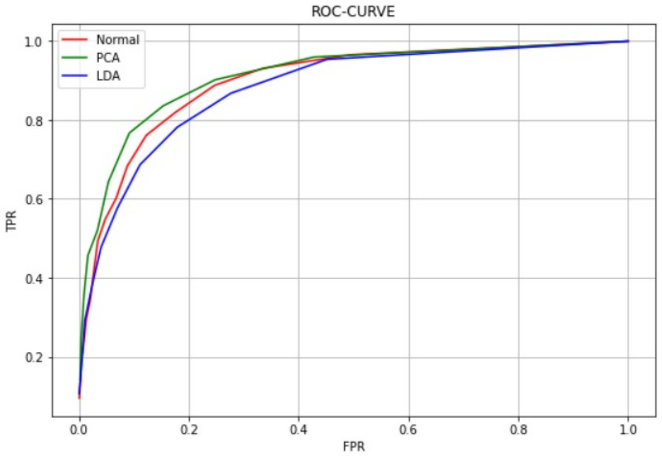


Fig. 5. ROC Curve of KNN applied on Image data, after PCA

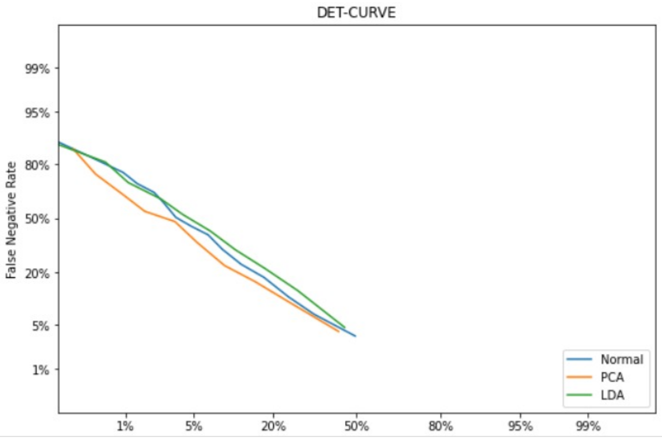


Fig. 6. DET Curve of KNN applied on Image data, after PCA

The following are the results for SVM Claasification; The Accuracy table has been plotted which consists of accuracy for each classifier and corresponding data given.
It can be observed that SVM yields the best result when trained on data after PCA is done.

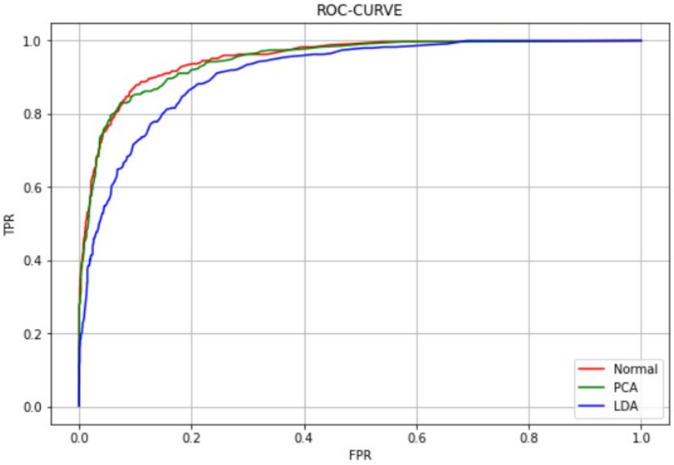


Fig. 7. ROC Curve of SVM applied on Image data

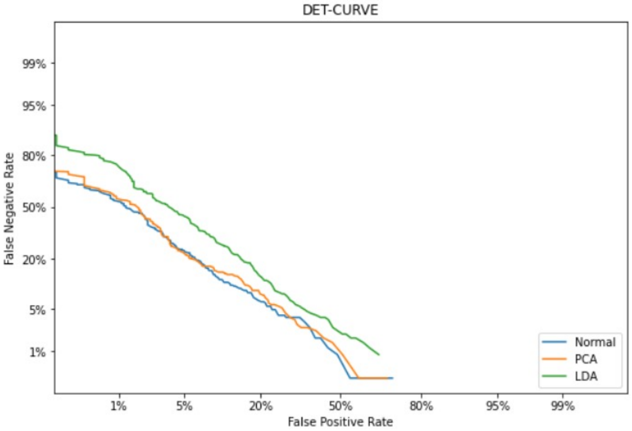


Fig. 8. DET Curve of SVM applied on Image data

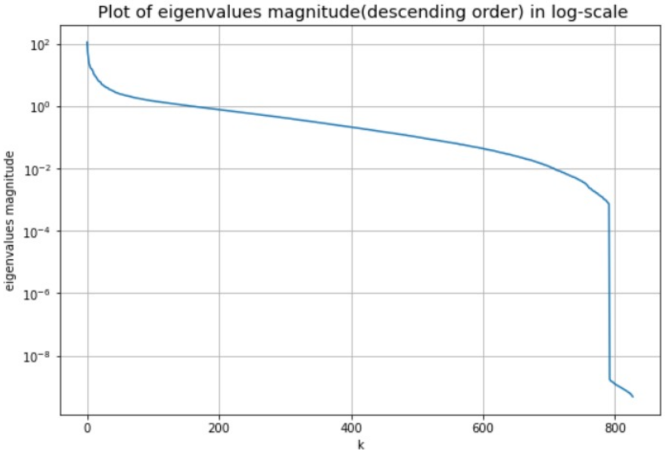


Fig. 9. Plot of Magnitude of Eigenvalues for PCA

	Normal	After PCA	After LDA
KNN	69.54	65.23	67.53
Logistic Regression	67.53	72.41	62.93
ANN	73.85	76.72	65.23
SVM	77.59	78.74	65.52

IV. Spoken Digits Data set

Speech Recordings of the numbers: **1, 3, 5, 9, z (Zero)** have been provided, from which the features also have been extracted.
These features are of varying length and hence we need to make them uniform before applying various models.
We achieve the same by re-sampling them using the scipy.signal.resample; all the vectors are transformed into Fourier domain and then re-sampled back to the mean length.

After all the data is transformed into uniform length, we initialise the KNN on the test data and train data to compute the results, which can be seen below:

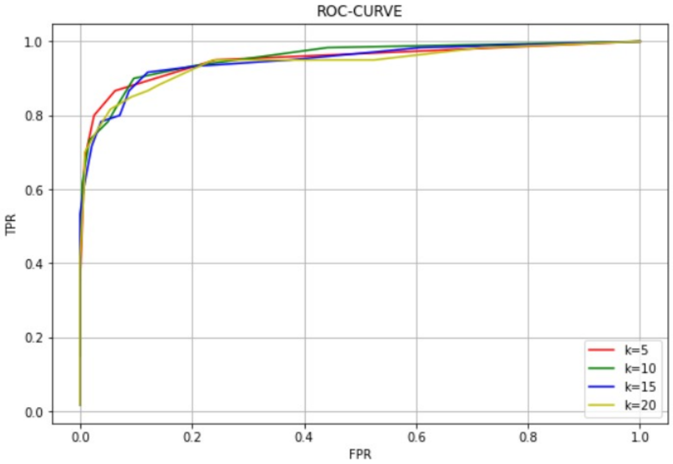


Fig. 10. ROC Curve of KNN on Spoken Digit data

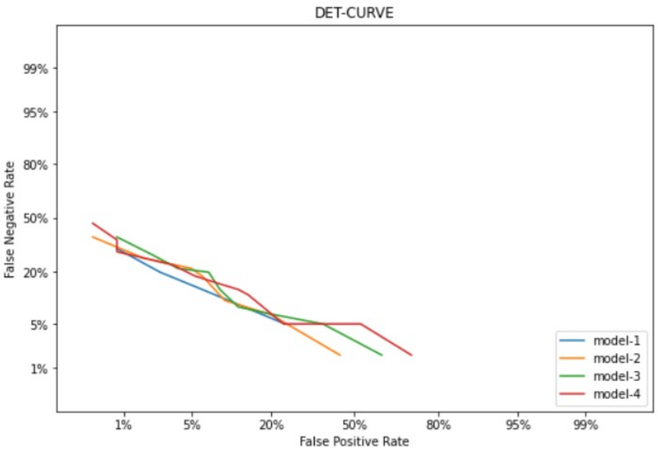


Fig. 11. DET Curve of KNN on Spoken Digit data

Now we apply PCA on the Spoken Digit data, i.e from the multiple feature data-set, we pick the top-k components and the test data is again classified; the underlying assumption is that we remove redundancy by picking only a select number of components and not all data.

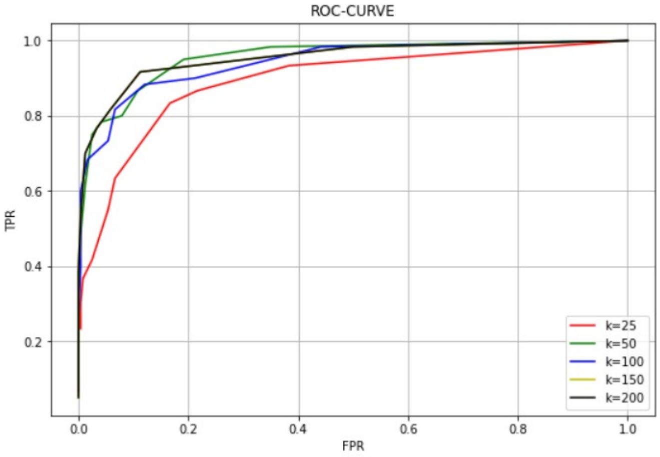


Fig. 12. ROC Curve of KNN on Spoken Digit data after PCA

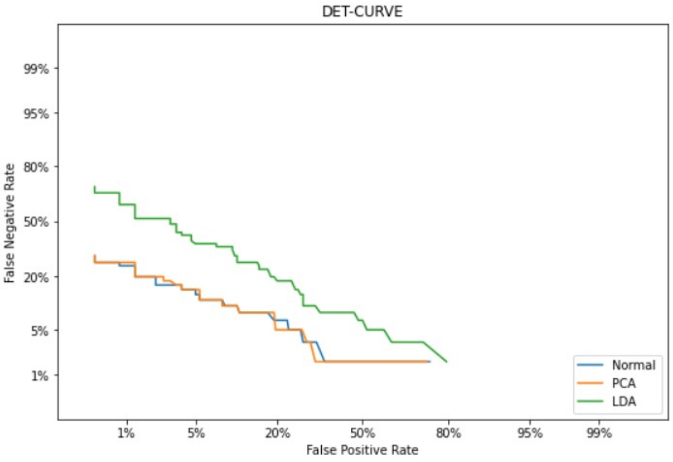


Fig. 13. DET Curve of KNN on Spoken Digit data after PCA

Accuracy Table:

	Normal	After PCA	After LDA
KNN	85	82	70
Logistic Regression	86.67	85	66.67
ANN	88.33	85	66.67
SVM	85	86.67	53.33

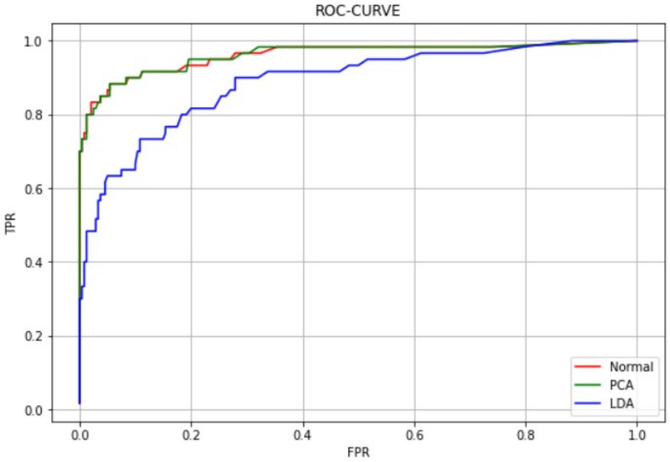


Fig. 14. ROC Curve of Logistic Regression

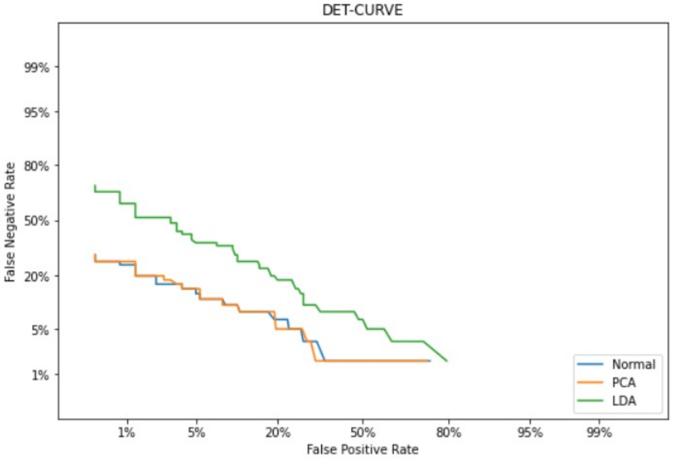


Fig. 15. DET Curve of Logistic Regression

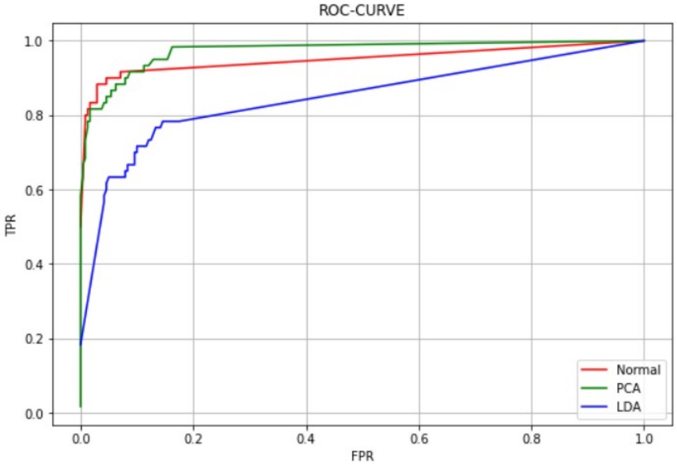


Fig. 16. ROC Curve of ANN applied on Spoken Digit data

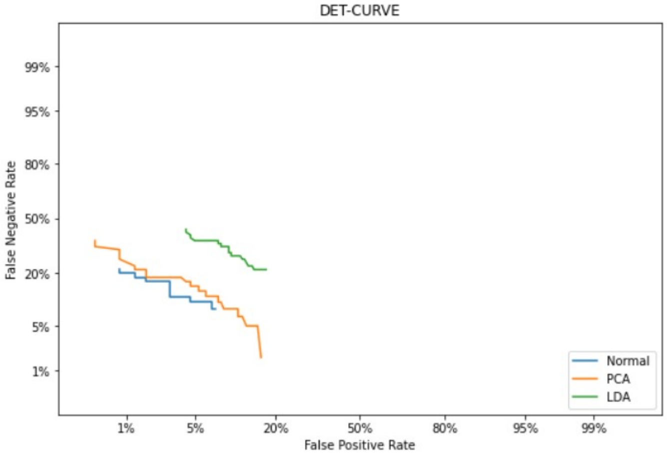


Fig. 17. DET Curve of ANN applied on Spoken Digit data

V. Handwritten Data Set

Hand written positions of the numbers: **a, ai, bA, cha, lA** have been provided, which will be appended to form one vector for each example
As performed before; These features are of varying length and hence we need to make them uniform before applying various models and hence We re-sample them using the scipy.signal.resample; all the vectors are transformed into Fourier domain and then re-sampled back to the mean length.

After all the data is transformed into uniform length, we initialise the KNN on the test data and train data to compute the results, which can be seen below:

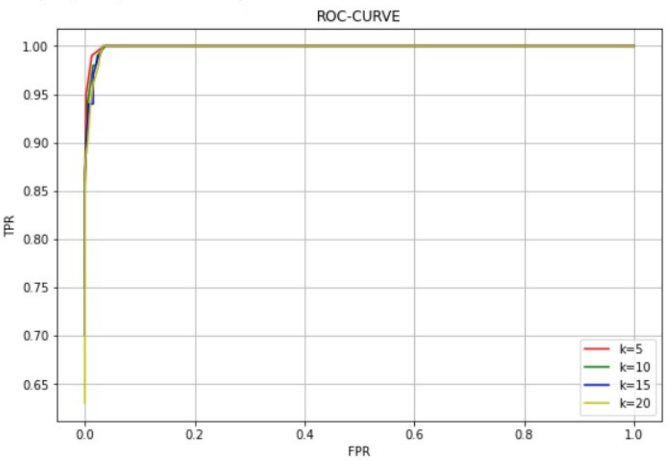


Fig. 18. ROC Curve of KNN on Handwritten data before PCA

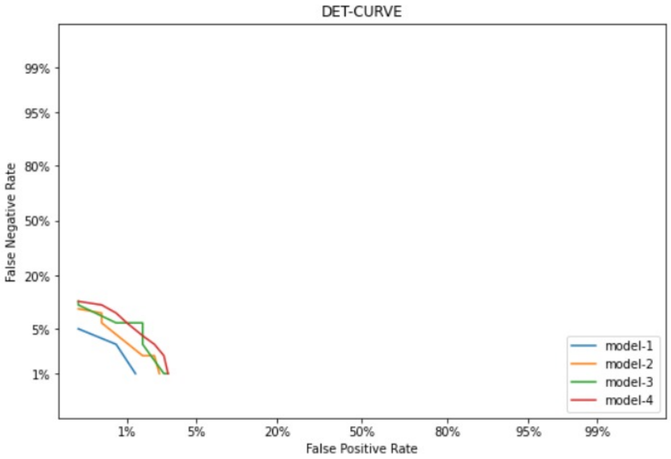


Fig. 19. DET Curve of KNN on Handwritten data before PCA

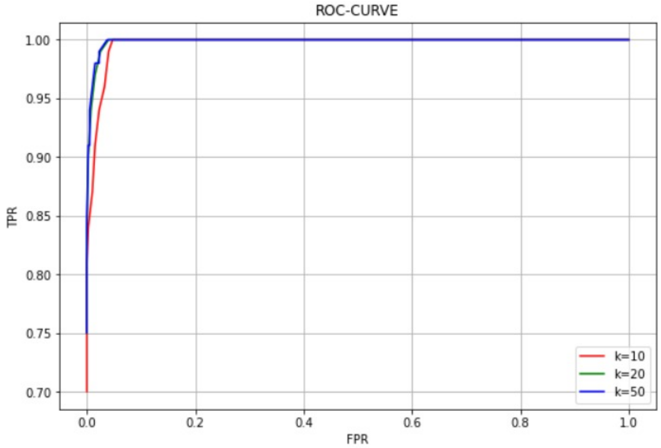


Fig. 20. ROC Curve of KNN on Handwritten data after PCA

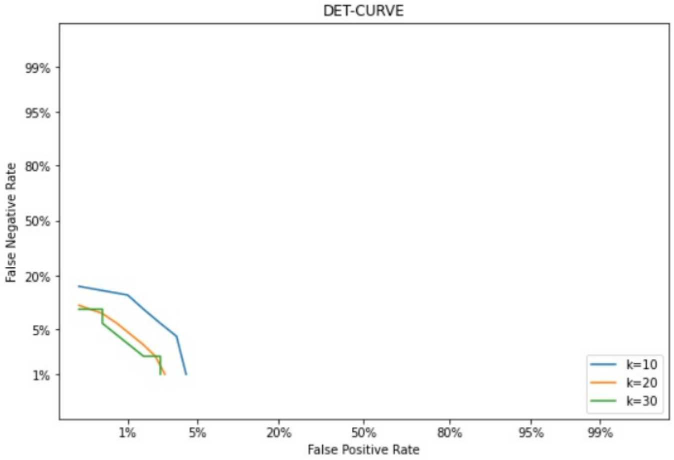


Fig. 21. DET Curve of KNN on Handwritten data after PCA

Accuracy Table:

	Normal	After PCA	After LDA
KNN	97	95	91
Logistic Regression	91	91	81
ANN	98	99	92
SVM	99	100	94

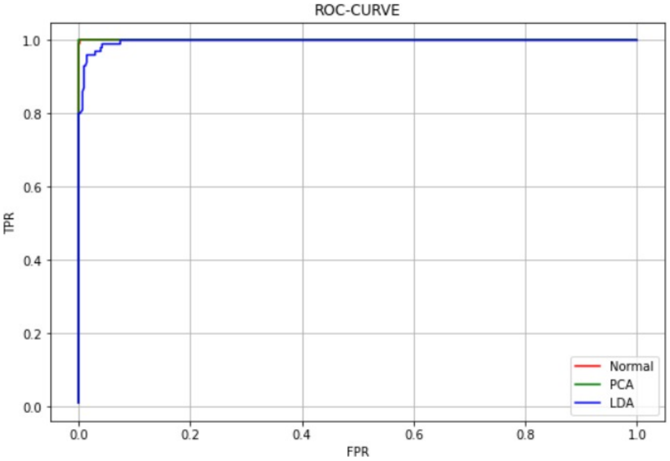


Fig. 22. ROC Curve of SVM applied on Handwritten data

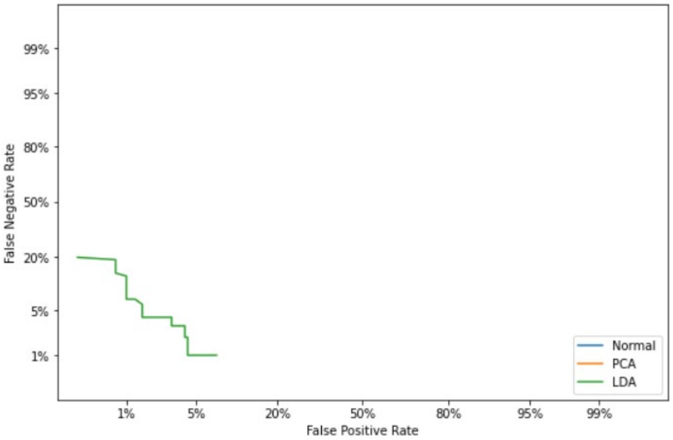


Fig. 23. DET Curve of SVM applied on Handwritten data

It can be observed that we attain fairly high accuracy, close to about 95-100% for the Handwritten data set.

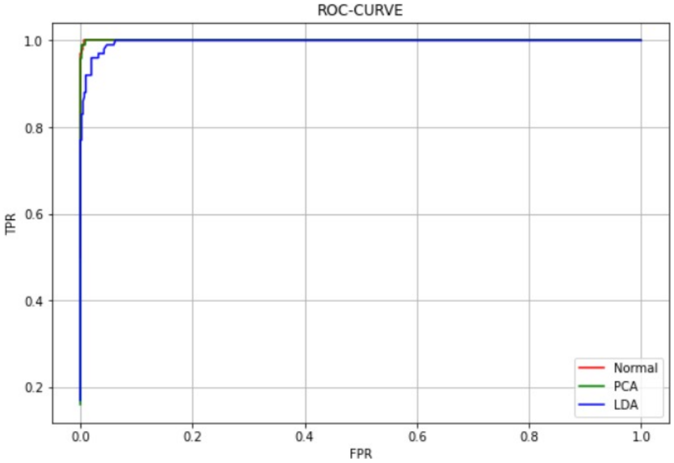


Fig. 24. ROC Curve of ANN applied on Handwritten data

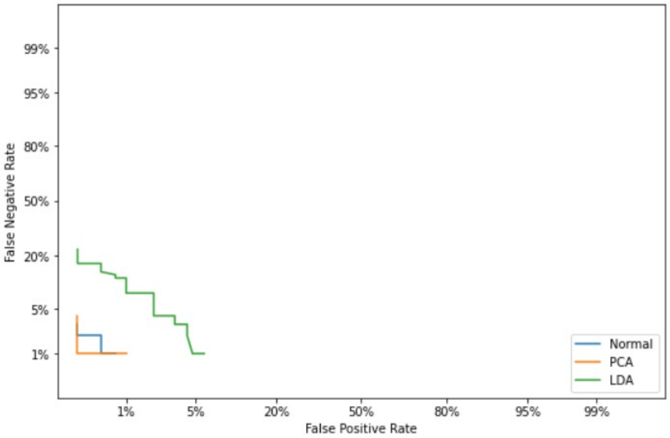


Fig. 25. DET Curve of ANN applied on Handwritten data